

基于融合数据自表示的离群点检测算法

高亚星,赵旭俊,曹栩阳

(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘要:数据自表示方法可以用于离群点检测,起到了放大数据间差异性和关联性的作用,但现有技术未能体现特征之间关联性对离群点检测的影响,因此无法用于高维数据。针对这个问题,提出了一种基于融合数据自表示的离群点检测算法,它可以有效地检测出高维数据中的离群点。首先,提出了一种基于特征关系的数据自表示方法,结合互信息与信息熵理论,度量高维数据特征间的关联性,并将其融于数据间的稀疏表示过程,体现了特征间和数据间的复杂关系。其次,提出了一种基于融合组间数据自表示的计算方法,采用点乘的方式将不同特征分组对应的自表示矩阵融为一体,形成全局数据自表示矩阵。最后,提出基于融合数据自表示的离群点检测算法,在全局数据自表示矩阵形成的有向加权图上,通过图随机游走检测离群点。实验结果表明,该算法在真实数据集和人工合成数据集上的检测性能均高于对比算法,证明该算法具有良好的泛化性和稳定性。

关键词:离群点检测;数据自表示;特征分组;信息熵;随机游走

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2023)12-0041-08

doi:10.3969/j.issn.1673-629X.2023.12.006

An Outlier Detection Algorithm Based on Fusion Data Self-representation

GAO Ya-xing, ZHAO Xu-jun, CAO Xu-yang

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: Data self-representation method can be used for outlier detection, which plays a role in magnifying the difference and correlation among data. However, the existing technologies fail to reflect the influence of correlation among features on outlier detection, so it cannot be used for high-dimensional data. To solve this problem, an outlier detection algorithm based on fusion data self-representation is proposed, which can effectively detect outliers in high-dimensional data. Firstly, a data self-representation method based on feature correlation is proposed, which combines mutual information and information entropy theory to measure the correlation among features of high-dimensional data, and integrates it into the sparse representation process among data, reflecting the complex relationship among features and data. Secondly, a calculation method based on the data self-representation among fusion groups is proposed. The self-representation matrix corresponding to different feature groups is integrated by point multiplication to form a global data self-representation matrix. Finally, an outlier detection algorithm based on fusion data self-representation is proposed. On the directed weighted graph formed by the global data self-representation matrix, outliers are detected by graph random walk. The experimental results show that the detection performance of the proposed algorithm on real datasets and synthetic datasets is higher than that of the comparison algorithm, which proves that the proposed algorithm has good generalization and stability.

Key words: outlier detection; data self-representation; feature grouping; information entropy; random walk

0 引言

离群数据是指与其他数据分布有显著不同的数据对象^[1]。离群点检测是通过分析离群数据的特征,从海量数据中挖掘异常信息和提取兴趣模式的一种方

法,已广泛地应用在欺诈检测^[2]、入侵检测^[3]、疾病检测^[4]、时间序列离群值检测^[5]等领域。

基于数据自表示的离群点检测^[6]是一种有效的方法,该方法通过构建数据点间的稀疏表示,放大了数

收稿日期:2023-02-08

修回日期:2023-06-09

基金项目:国家自然科学基金(61572343);国防科技重点实验室基金项目资助(JSY6142219202114);山西省应用基础研究计划项目(20210302123223,202103021224275)

作者简介:高亚星(1998-),女,硕士研究生,研究方向为数据挖掘;通信作者:赵旭俊(1976-),男,博士,教授,研究方向为数据挖掘与并行计算。

据间的差异性,在低维离群点检测上有较好的性能。但该方法仍存在一定的局限性,导致其在高维数据上的离群点检测效果不佳。随着数据量和数据维度的急速攀升,解决该问题是十分必要的。

基于数据自表示的离群点检测方法通过研究低维数据间的关联性与稀疏性,并构建表示关系,在低维全局离群点检测上获得了较好的效果。但该方法在处理高维数据时存在以下两点问题:第一,高维数据相较于低维数据会更加离散,使得该方法更难得出较为准确的数据表示,从而影响高维离群点检测的效果;第二,该方法并未考虑特征间的关联性对数据相互表示过程的影响,在处理低维数据时,这种局限性对检测结果影响并不大,然而高维数据中特征间的复杂关系不容忽视,这些特征关系对离群点的检测有较大影响,该方法的局限性被进一步放大。

为了解决以上问题,该文提出了一种基于融合数据自表示的离群点检测算法。首先,使用文献[7]提出的特征分组方法对数据集按照相关特征进行分组,达到数据约简的目的。其次,提出一种基于特征关系的数据自表示方法,在每个特征分组内,运用信息理论度量特征间的关联性,构建了包含特征和数据双重信息的数据自表示矩阵。然后,提出一种基于融合组间数据自表示的计算方法,将不同特征分组和组中心特征集对应的数据自表示矩阵相融合,采用矩阵点乘和均值计算得出全局自表示矩阵,其包含了丰富的全局数据特征和数据信息。最后,提出了基于融合数据自表示的离群点检测算法,该算法在全局数据自表示矩阵上构建了有向加权图,并通过在该图上随机游走来检测离群点。在多个真实数据集和人工合成数据集上的实验结果表明,该算法具有良好的泛化性和稳定性。

1 相关工作

目前,传统离群点检测算法主要分为四大类,分别是基于密度的方法^[8]、基于距离的方法^[9]、基于近邻的方法^[10]和基于聚类的方法^[11]。随着高维数据的出现,数据变得更加稀疏^[6],全局离群值的挖掘受到更多不相关特征的影响,使得传统的离群点检测算法时间复杂度较高且检测效果不佳。

提出基于子空间的方法^[12-13]的目的是降低高维数据出现带来的影响,此类方法将所有数据点映射到一个或多个稀疏和低维子空间中进一步检测离群点。文献[12]提出了一种基于特征异常值相关分析的局部离群值检测算法,由于其删除了不相关数据点和数据特征,虽更适用于高维数据,但难以保证精度。文献[13]提出利用随机哈希方法对子空间内数据点进行评分。然而随着数据集规模的增长,它的复杂性

迅速增加。由于基于子空间的方法面向数据点进行划分,较少地考虑数据特征间的关系,仍具有不足之处。基于特征分组的离群点检测算法有效弥补了上述缺陷。文献[7]提出基于互信息和信息熵的特征分组方法,细化了特征间相关性度量,但是该方法在检测全局离群点时忽略了不同特征分组间仍存在的关系,影响了检测效果。

近些年提出了基于数据自表示的离群点检测算法^[6],此类方法时间复杂度较低,提升了高维离群点检测效率,但仍存在一些问题。文献[6]提出了基于数据自表示随机游走离群点检测方法,构建了数据点之间的稀疏表示,得到了较好的检测效果,但由于未考虑数据特征间的相关性,使其检测结果会丢失许多特征信息。文献[14]提出了一种基于数据结构信息度量的双核参数方法,获得了更加准确的数据表示结果,但此方法存在参数依赖。文献[15]提出了基于超像素的张量低秩分解,将其投影到低维空间检测离群数据。但该方法受到初始字典的约束,同时在投影过程中会丢失部分信息。文献[16]使用核范数而不是 Schatten-p 范数来获得更准确的数据低秩表示。以上几种方法虽有效降低了时间复杂度,但均存在参数依赖问题,同时处理高维数据时都未考虑数据特征间存在的复杂关系,从而影响了离群点检测效果。

综上所述,基于特征分组的离群点检测方法虽降低了检测维度,但在全局离群点的检测中会丢失特征分组之间的相关性信息,影响了离群点检测性能;基于数据自表示的离群点检测方法在保证检测结果准确性的同时,有效地降低了时间复杂度,但现有的数据自表示方法同样丢失高维特征包含的有效信息,离群点检测未能更全面和深入,从而影响最终的检测效果。

2 基于特征关系的数据自表示方法

该文提出的基于特征关系的数据自表示方法,结合互信息与条件熵理论来度量特征间的关联性,并将其融于数据自表示过程。互信息用于量化两个随机变量之间的依赖程度。其值越高代表两个变量之间相关性越高,一个变量在一定程度上可以被另一个变量所取代。条件熵用于描述已知在一个随机变量的条件下,另一个随机变量的不确定性。

设 $D = \{d_1, d_2, \dots, d_n\}$ 为任意数据集, $F = \{f_1, f_2, \dots, f_l\}$ 为该数据集中所有特征的集合,其中: d_n 表示第 n 个数据点, f_l 表示第 l 个特征, n 和 l 分别表示数据点和数据特征的个数。 $G = \{F_{G1}, F_{G2}, \dots, F_{Gm}\}$ 为算法 FG^[7] 得出的特征分组集合,其中 F_{Gm} 表示第 m 个特征分组, m 表示总分组数。

任一特征分组对应的数据自表示矩阵记为 $R =$

$\{r_1, r_2, \dots, r_n\}$, 以 $n \times n$ 方阵的形式记录了这种表示关系, 其主对角线上的值 (r_{ii}) 为 0。 \mathbf{R} 由使 $R(r_i)$ 值最小时的 r_i 构成, $R(r_i)$ 的计算公式如下:

$$R(r_i) = \sum_{j=1}^c \|r_i\|_2^2 + \frac{1-\delta}{2} \|r_i\|_F^2 + \frac{\omega}{2} \|F_{G_j} - F_G \cdot r_i\|_2^2 \quad (1)$$

其中, c 表示该特征分组内的特征总数, r_i 表示矩阵 \mathbf{R} 中的第 i 列, F_G 表示该特征分组集合, F_{G_j} 表示第 j 个特征。系数 δ 和 ω 由公式(2)和公式(5)定义。

δ 用于均衡地度量特征之间的相似性与相关性, 将高斯核函数用于计算特征间的距离以表示相似性, 同时互信息与条件熵用于度量特征间的概率关系以表示相关性, 二者结合使得特征间的复杂关系得以深入且全面的表达。 δ 可由如下公式得出:

$$\delta = \exp\left(-\frac{\|F_{G_i} - F_{G_j}\|_2}{\sigma_{F_{G_i}}^2 + \sigma_{F_{G_j}}^2} - \frac{I(F_{G_i}, F_{G_j})}{H(F_{G_i} | F_{G_j})}\right) \quad (2)$$

其中, F_{G_i} 表示第 i 个特征, F_{G_j} 表示第 j 个特征, 其余符号同公式(1)中定义。 $I(F_{G_i}, F_{G_j})$ 表示 F_{G_i} 和 F_{G_j} 之间的互信息, 可由公式(3)得出。 $H(F_{G_i} | F_{G_j})$ 表示 F_{G_j} 在 F_{G_i} 条件下的条件熵, 可由公式(4)得出。

$$I(F_{G_i}, F_{G_j}) = \sum_{f_{G_i} \in F_{G_i}} \sum_{f_{G_j} \in F_{G_j}} p(f_{G_i}, f_{G_j}) \log_2 \frac{p(f_{G_i}, f_{G_j})}{p(f_{G_i}) \times p(f_{G_j})} \quad (3)$$

$$H(F_{G_i} | F_{G_j}) = - \sum_{f_{G_i} \in F_{G_i}} \sum_{f_{G_j} \in F_{G_j}} p(f_{G_i}, f_{G_j}) \log_2 p(f_{G_i} | f_{G_j}) \quad (4)$$

其中, f_{G_i} 和 f_{G_j} 表示第 i 和第 j 个特征的值。

由于具有强相关性的特征会被划分至同一分组, 出现特征冗余现象, 一定程度上浪费了计算资源。 ω 用于度量分组内特征间的冗余程度, 提升了数据自表示结果的有效性。 ω 可由如下公式得出:

$$\omega = \frac{\sum_{F_{G_i}, F_{G_j} \in F_G} I(F_{G_i}, F_{G_j})}{|F_G|^2} \quad (5)$$

此节提出的基于特征关系的数据自表示方法, 通过度量特征间存在的复杂关系, 弥补了现有算法的局限性。该方法构建了数据点之间的稀疏线性组合, 正常点仅由正常点表示, 而离群点可以由正常点和离群点共同表示, 并通过 δ , ω 和范数度量特征和数据的相关性, 以达到将特征信息与数据信息相融合的效果。

3 融合组间数据自表示与离群点检测

根据上一节的方法, 每个特征分组生成了对应的数据自表示矩阵, 其中包含各分组的特征和数据分布信息, 这种信息是局部的。由于不存在某个特征同时属于两个或更多分组的情况, 导致数据自表示矩阵也

是唯一对应的, 使得原始数据被简化, 造成不同矩阵中包含的离群数据信息存在差异。若直接在其基础上检测离群点所得到的结果不足以涵盖全局信息, 又因为不同特征分组间仍然存在关联性, 这些信息对于准确地挖掘全局离群点有帮助, 进一步研究融合组间关联性将提升检测结果有效性。

3.1 基于融合组间数据自表示的计算方法

通过融合特征组间关联性构建的全局数据自表示矩阵 (\mathbf{RT}) 可由如下公式得出:

$$\mathbf{RT} = \frac{1}{2}(\mathbf{RC} + \mathbf{R}_1 \cdot \mathbf{R}_2 \cdot \dots \cdot \mathbf{R}_m) \quad (6)$$

其中, FG 算法依据其定义的相关性度量方式, 得到各组中心特征, 具有与组内其余特征均强相关的特点。FG 算法得出的组中心特征集表示为 $\mathbf{FC} = \{\mathbf{FC}_1, \mathbf{FC}_2, \dots, \mathbf{FC}_m\}$, \mathbf{RC} 表示 \mathbf{FC} 上的数据自表示矩阵, $\mathbf{R}_1 \cdot \mathbf{R}_2 \cdot \dots \cdot \mathbf{R}_m$ 分别表示 m 个特征分组对应的数据自表示矩阵。公式(6)中, 通过构建中心特征分组对应的数据自表示矩阵 (\mathbf{RC}), 将不同组中心融合为一体, 使得每个特征分组的主要信息被集中表达, 均衡了组间差异性。 $\mathbf{R}_1 \cdot \mathbf{R}_2 \cdot \dots \cdot \mathbf{R}_m$ 表示通过点乘各自表示矩阵, 使得正常点与离群程度较大的点相互表示值与其他表示关系之间的差值增大, 某些可能是离群点的数据与正常点之间的关系更弱。该文提出的基于融合组间数据自表示的计算方法, 通过融合各特征分组的数据自表示矩阵, 加入组间特征信息, 构建全局矩阵, 为后续离群点检测提供了良好的基础。

3.2 基于融合数据自表示的离群点检测算法

该文提出基于融合数据自表示的离群点检测算法, 通过在构造的加权有向图上随机游走来筛选离群点。该加权有向图 (S) 由全局数据自表示矩阵 (\mathbf{RT}) 构造得出, S 中有向边为 \mathbf{RT} 中数据点间的相互指向, 边权值表示数据点之间的相关性, 同时也表示两点之间的转移概率。 S 的边权值可由公式(7)得出:

$$p_{xy} = w_{xy} = \frac{r_{xy}}{\sum_{y=1}^n |r_{xy}|} \quad (7)$$

其中, p_{xy} 表示数据点 x 到数据点 y 的转移概率, w_{xy} 表示图 S 上数据点 x 与数据点 y 之间的边权值, 与 p_{xy} 值相同。 r_{xy} 表示 \mathbf{RT} 中第 x 行第 y 列的值。

阻尼因子 (η) 在文献[14]中被用于马尔可夫转移概率的计算。公式(8)描述了 S 上不同时间的状态转移分布, 随机选取初始点。

$$S_{i \in l}^{(t+1)} = \eta \cdot \mathbf{O} + (1 - \eta) \cdot S_{i \in l}^{(t)} \cdot \mathbf{P} \quad (8)$$

其中, $S^{(t)}$ 表示 t 时刻的状态转移分布, \mathbf{P} 表示转移概率矩阵, \mathbf{O} 表示 $1 \times n$ 的矩阵, n 表示数据点个数。

基于稳定后的状态转移分布, 通过公式(9)得出

每个数据点的异常因子(od),其用于描述数据点的离群程度。正常点被转移到的概率较高,而离群点被转移到的概率较低,取异常因子最小的 h 个点作为最终的全局离群点。

$$od_x = 1 - \sum_{y=1}^n S_{xy}^{(t)} \quad (9)$$

其中, od_x 表示第 x 个数据点的异常因子, $S_{xy}^{(t)}$ 表示状态转移稳定后第 x 行第 y 列的值。

综上所述,基于融合数据自表示的离群点检测算法基本思想为:首先在 FG 算法的特征分组结果基础上,使用公式(1)构建每个分组和组中心特征集对应的数据自表示矩阵;然后,使用基于融合组间数据自表示的计算方法得出全局自表示;最后,在加权有向图上使用马尔可夫随机游走检测离群点。其算法描述如下:

算法 1: MDSR (An outlier detection algorithm based on merge data self-representation)

输入:FG 算法得出的特征分组: $G = \{F_{G1}, F_{G2}, \dots, F_{Gm}\}$, 组中心特征集为 $FC = \{FC_1, FC_2, \dots, FC_m\}$

输出:异常因子(od)

1. 根据公式(1),得出 G 中每个特征分组和 FC 对应的数据自表示矩阵(R)
2. 根据公式(6),构建全局数据自表示矩阵(RT)
3. 根据公式(7),使用 RT 计算出状态转移矩阵(P)
4. $\eta \leftarrow 0.1, S^{(0)} \leftarrow \{1/n, 1/n, \dots, 1/n\}$
5. 根据公式(8),得出稳定的状态转移分布($S^{(t)}$)
6. 根据公式(9),计算出每个数据点的异常因子(od),并取其值最小的 h 个点作为最终的全局离群点。

4 实验分析

4.1 实验设置及环境

为了验证 MDSR 算法的有效性,在人工数据集和真实数据集上共同实验,使用 LOF^[17], SOD^[18], ROD^[19], COPOD^[20] 和 R-GRAPH^[6] 作为对比算法,采用 ROC 曲线下面积(AUC)和运行时间作为评价指标。真实数据集包含 3 个 UCI Machine Learning Repository (UCI) 数据集和 6 个 Outlier Detection DataSets (ODDS) 数据集,详见表 1。

表 1 真实数据集

数据集	数据量	特征数	来源
Breast-Cancer	699	10	UCI
Breast-Cancer (Diagnostic)	569	32	UCI
Ann-thyroid	3 772	6	UCI
Ionosphere	351	33	ODDS
Pendigits	6 870	16	ODDS
Optdigits	5 216	64	ODDS
Musk	3 062	166	ODDS

续表 1

数据集	数据量	特征数	来源
Arrhythmia	452	274	ODDS
Speech	3 686	400	ODDS

人工数据集来源于文献[21],使用文献[8]中对人工数据集的处理方法,得到本实验中的 28 个数据集,分别为:18 个数据规模一定的数据集,其中特征数量包含 20,30,40,50,75 和 100,每个特征数量中包含 3 个不同的数据集;与 10 个特征数量一定的数据集,其中特征数为 20,数据量从 1 000 到 10 000,详见表 2。该文基于 Matlab 平台,AMD Ryzen 7 4800H 2.90 GHz CPU 和 RTX 2060 GPU 实现了 MDSR 算法。

表 2 人工数据集

数据集	数据量	特征数	来源
D20_a, D20_b, D20_c	1 000	20	人工
D30_a, D30_b, D30_c	1 000	30	人工
D40_a, D40_b, D40_c	1 000	40	人工
D50_a, D50_b, D50_c	1 000	50	人工
D75_a, D75_b, D75_c	1 000	75	人工
D100_a, D100_b, D100_c	1 000	100	人工
D20_1, D20_2, ..., D20_10	1 000, 2 000, ..., 10 000	20	人工

4.2 人工数据集实验结果分析

6 种算法在 6 个不同维度人工数据集上的 AUC 如图 1 所示,具体 AUC 数值如表 3 所示,取每维度中的 3 个数据集上最大的 AUC 作为实验结果。依据参考文献[8,22],LOF 和 SOD 算法中参数 $n_neighbors$ 的值取 5 时其算法模型达到最优。

从图 1 和表 3 可见,对于所有维度的人工数据集, MDSR 算法的 AUC 值明显比 LOF, SOD, ROD 和 COPOD 算法的高,较好地完成了离群点检测任务。

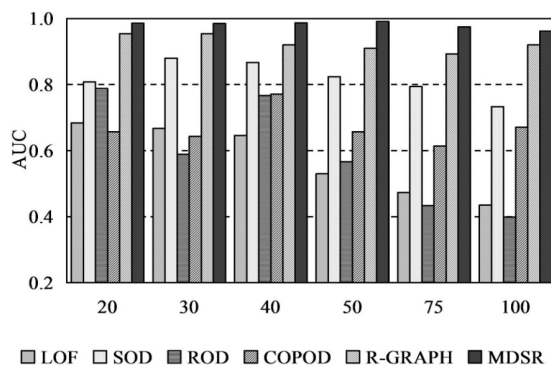


图 1 人工数据集上的 AUC 对比

LOF 和 ROD 算法性能随着数据维度的增加而大幅下降,造成这种现象的原因是,即使 ROD 算法提出了基于全维 3D 子空间的方法,它们仍然都在全局数

据中直接检测离群值,当处理高维数据时,对于 LOF 和 ROD 算法将更难检测离群值。COPOD 算法 AUC 值则在小范围内浮动,是因为它更多受到数据分布的影响,而非特征的数量。SOD 算法构造的子空间中包含冗余特征,数据特征越多,冗余程度越大,导致 SOD 同样难以有效处理高维数据。提出的 MDSR 算法的

AUC 值比 R-GRAPH 的略高,是因为 MDSR 不仅考虑了特征冗余问题还将特征间复杂的关联性融于数据自表示的过程中,而 R-GRAPH 在其数据自表示过程中并未关注这些问题。因此 MDSR 算法性能不会随着特征数的增加而显著降低,从而保证其面对高维数据的离群点检测任务时,仍可以得到较准确的检测结果。

表 3 人工数据集上的 AUC 数值

特征数	LOF	SOD	ROD	COP-OD	R-GRAPH	MDSR
20	0.684	0.808	0.788	0.657	0.954	0.986
30	0.667	0.880	0.589	0.643	0.954	0.985
40	0.646	0.867	0.767	0.771	0.920	0.987
50	0.530	0.824	0.566	0.657	0.910	0.991
75	0.473	0.794	0.433	0.614	0.893	0.975
100	0.435	0.733	0.399	0.671	0.920	0.962

离群点检测算法的运行时间是衡量其检测效率的一个重要评价指标。各算法在人工数据集上的运行时间对比如表 4 所示。

表 4 不同特征数量人工数据集上的运行时间对比

特征数	LOF	SOD	ROD	COP-OD	R-GRAPH	MDSR
20	0.194	4.090	50.410	0.031	0.092	0.910
30	0.187	4.060	166.890	0.043	0.102	1.863
40	0.190	4.150	393.710	0.049	0.194	3.372
50	0.200	4.280	440.200	0.051	0.197	5.244
75	0.218	4.200	2 771.08	0.092	0.226	11.923
100	0.208	4.202	6 749.65	0.123	0.248	21.161

从表 4 可见,MDSR 算法在前 3 个数据集上的运行时间比 SOD 算法的少,在全部数据集上的运行时间比 ROD 算法的少。对于 SOD 算法,选择相关子空间的过程占用了运行时间。对于 ROD 算法,构建全局三维子空间的过程有较大的时间开销。由于 COPOD 算法是基于连接的,在数据集规模为 1 000 的前提下,其

运行时间浮动较小。提出的 MDSR 算法深入挖掘特征关系,随着数据特征的增加,这种复杂关系更加难以计算,导致该方法在运行时间上略有增加。为分析数据规模对 MDSR 算法运行时间的影响,在数据集 D20_1, D20_2, …, D20_10 上进一步对比,结果如表 5 所示。

表 5 不同数据量人工数据集上的运行时间对比

数据量	LOF	SOD	ROD	COP-OD	R-GRAPH	MDSR
1 000	0.194	4.091	50.410	0.031	0.092	0.910
2 000	0.343	7.145	88.519	0.043	0.251	1.071
3 000	0.703	12.140	136.160	0.063	0.439	1.152
4 000	1.184	19.090	182.150	0.078	0.584	1.404
5 000	1.787	27.910	225.740	0.094	0.850	1.603
6 000	2.266	40.210	276.802	0.125	1.012	1.739
7 000	2.983	51.927	290.750	0.157	1.273	2.030
8 000	3.439	67.280	326.690	0.171	1.437	2.257
9 000	4.455	83.064	393.860	0.187	1.748	2.583
10 000	5.115	96.379	449.440	0.210	1.878	2.892

从表 5 可见,实验中所有算法的运行时间都随数据量增加而线性增加,但 MDSR 算法的运行时间仍保

持在较低值,且远小于 SOD 和 ROD 算法的运行时间,更直观的对比如图 2 所示。

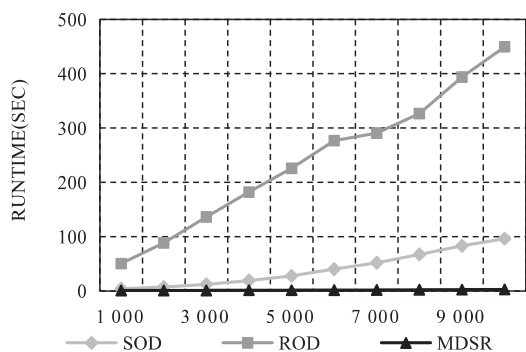


图 2 不同数据规模下 SOD, ROD 和 MDSR 算法运行时间对比

由于 MDSR 算法考虑了特征间关联性,使其运行时间比 COPOD 和 R-GRAPH 算法的较长。但是结合

表 3 中 AUC 数值和表 4、表 5 中运行时间可以得出,该文提出的算法运行时间虽不是最短,但仍取得了最好的检测结果,同时结合数据量大小和数据维度对 MDSR 算法性能的影响,表明该算法适合用于高维离群点检测任务且总体优于对比算法。

4.3 真实数据集实验结果分析

为进一步验证 MDSR 算法的有效性,将其与 LOF, SOD, ROD, COPOD 和 R-GRAPH 在 9 个真实数据集上进行了对比实验,且设置了 3 种离群点占比,分别为 2%, 5% 和 10%。由于 Pendigits, Optdigits, Musk 和 Speech 数据集中离群点占比不足 5%,故只以 2% 的离群值作为实验设置。3 种离群点占比下不同算法在真实数据集上的 AUC 对比如表 6 所示。

表 6 真实数据集上的 AUC 对比

数据集	离群占比	LOF	SOD	ROD	COPOD	R-GRAPH	MDSR
Breast Cancer	2%	0.849	0.888	0.616	0.923	0.933	0.968
	5%	0.784	0.819	0.526	0.868	0.928	0.957
	10%	0.679	0.780	0.559	0.744	0.933	0.970
Breast Cancer (Diagnostic)	2%	0.851	0.926	0.734	0.918	0.957	0.971
	5%	0.877	0.905	0.729	0.917	0.966	0.975
	10%	0.860	0.875	0.644	0.844	0.946	0.975
Ann-thyroid	2%	0.945	0.975	0.543	0.728	0.946	0.952
	5%	0.952	0.973	0.695	0.650	0.950	0.951
	10%	0.915	0.960	0.656	0.627	0.933	0.955
Arrhythmia	2%	0.729	0.660	—	0.918	0.923	0.940
	5%	0.609	0.736	—	0.927	0.943	0.968
	10%	0.612	0.839	—	0.884	0.942	0.953
Ionosphere	2%	0.796	0.936	0.745	0.933	0.950	0.958
	5%	0.767	0.928	0.668	0.924	0.925	0.960
	10%	0.791	0.958	0.611	0.882	0.942	0.960
Pen digits		0.919	0.904	0.589	0.751	0.942	0.969
Opt digits	2%	0.81	0.951	0.489	0.766	0.912	0.974
Musk		0.537	0.763	—	0.776	0.914	0.936
Speech		0.533	0.677	—	0.722	0.856	0.905

从表 6 可见, MDSR 算法具有较好的检测结果。结合表 1, MDSR 算法无论在较低维数据集 Breast Cancer 和 Ionosphere 上,还是在较高维数据集 Musk, Arrhythmia 和 Speech 上, AUC 值都明显比 LOF, SOD 和 ROD 算法的高。由于 LOF 算法原理较为直接和单一,导致其在高维数据集上无法保证检测效果。MDSR 算法考虑了特征间关系和数据信息的融合,优于仅研究特征信息的 SOD 算法,使得 MDSR 算法的检测效果更突出。对于 ROD 算法,从表 6 可见其

AUC 值大于 0.7, 只出现在低维数据集 Breast-Cancer (Diagnostic) 和 Ionosphere 上。由于 ROD 在检测过程中旋转其构建的 3D 子空间,此过程需要占用的内存随数据维度增加而变大,而本设备不足以支撑 ROD 算法检测高维数据集 Arrhythmia 和 Speech。MDSR 算法的 AUC 值相比 R-GRAPH 的略优,原因是 R-GRAPH 只通过研究数据间相互表示关系检测离群点,存在一定的局限性。对于基于数据连接的离群点检测算法 (COPOD), 当数据量小于 1 000 时, 它的检测 AUC 可

以达到 0.9 以上,但数据量一旦超过 3 000,其检测 AUC 下降幅度超过了 0.3。

结合以上分析可以得出,MDSR 算法更加全面和深入地挖掘了数据和特征信息,具有较好的稳定性和

泛化性。

各算法在真实数据集上的运行时间对比如表 7 所示。

表 7 真实数据集上的运行时间对比

数据集	离群占比	LOF	SOD	ROD	COPOD	R-GRAPH	MDSR
Breast Cancer	2%	0.108	3.340	2.156	0.007	0.016	0.118
	5%	0.101	3.310	2.116	0.007	0.020	0.121
	10%	0.106	3.260	2.150	0.007	0.017	0.120
Breast Cancer (Diagnostic)	2%	0.148	3.820	135.3	0.025	0.051	0.241
	5%	0.101	3.290	136.0	0.025	0.053	0.240
	10%	0.100	3.210	134.8	0.024	0.050	0.240
Ann- thyroid	2%	0.172	16.06	7.650	0.025	0.040	0.173
	5%	0.172	15.50	6.050	0.024	0.064	0.170
	10%	0.157	14.62	6.011	0.025	0.027	0.171
Arrhy- thmia	2%	0.178	6.500	—	0.175	0.048	26.027
	5%	0.182	6.433	—	0.172	0.045	26.027
	10%	0.18	6.452	—	0.171	0.044	26.029
Iono- sphere	2%	0.17	3.230	102.8	0.022	0.053	0.469
	5%	0.163	3.220	103.7	0.022	0.065	0.465
	10%	0.161	3.129	103.2	0.022	0.059	0.469
Pen digits	2%	3.07	58.20	226.7	0.122	0.019	0.629
Opt digits		2.04	36.40	1 900	0.217	0.031	21.404
Musk		0.893	12.11	—	0.476	0.032	53.863
Speech		0.491	17.28	—	1.485	0.116	86.227

从表 7 可见,MDSR 算法的运行时间在所有数据集上少于 SOD 和 ROD 算法的运行时间,而只在低维数据集上少于 LOF 算法的运行时间。SOD 算法将时间消耗在寻找子空间上,ROD 算法则在构造三维子空间和旋转检测上花费较多时间,导致其运行时间随着数据特征的增多大幅上升。虽然 COPOD 算法运行时间较短,但该算法并不能保证离群点检测效果。MDSR 算法需要较多的时间用于计算特征间关联性,但为了能在高维数据中更有效地挖掘出离群点,可以忽略这部分时间开销。

5 结束语

通过均衡地度量特征与数据关联,该文提出一种基于融合数据自表示的离群点检测算法,充分挖掘并体现了数据特征间和数据间的复杂关系,有效改善了高维离群点检测性能,适用于高维数据的离群点检测任务。算法度量特征间相关性的过程较为费时,下一步主要研究工作是采取一些方法来优化该过程,已在保证检测有效性的基础上降低其时间消耗。

参考文献:

[1] HAWKINS D M. Identification of outliers[M]. London: Chapman and Hall,1980.

[2] CHADYŠAS V,BUGAJEV A,KRIAUIZENĖ R,et al. Outlier analysis for telecom fraud detection[C]//Proceedings of international baltic conference on digital business and intelligent systems. [s. l.]:Springer,2022:219-231.

[3] ZHANG X,HAN Y,XU W,et al. HOBA:a novel feature engineering methodology for credit card fraud detection with a deep learning architecture[J]. Information Sciences,2021, 557:302-316.

[4] ERTEN C,HOUDJEDJ A,KAZAN H. Ranking cancer drivers via betweenness-based outlier detection and random walks[J]. BMC Bioinformatics,2021,22(1):1-16.

[5] LAI K H,ZHA D,WANG G,et al. Tods:an automated time series outlier detection system [C]//Proceedings of the AAAI conference on artificial intelligence. [s. l.]:AAAI, 2021:16060-16062.

[6] YOU C,ROBINSON D P,VIDAL R. Provable self-repre-

- sensation based outlier detection in a union of subspaces [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii; IEEE, 2017; 3395 – 3404.
- [7] LI J, ZHANG J, PANG N, et al. Weighted outlier detection of high-dimensional categorical data using feature grouping [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 50(11): 4295–4308.
- [8] RIAHI-MADVAR M, AZIRANI A A, NASERSHARIF B, et al. A new density-based subspace selection method using mutual information for high dimensional outlier detection [J]. Knowledge-Based Systems, 2021, 216: 106733.
- [9] 俞庆英, 罗永龙, 陈付龙, 等. 一种保护私有信息的空间离群检测方法[J]. 计算机工程, 2017, 43(3): 163–171.
- [10] KRIEGEL H P, KRÖGER P, SCHUBERT E, et al. Outlier detection in axis-parallel subspaces of high dimensional data [C]//Proceedings of pacific-Asia conference on knowledge discovery and data mining. Berlin; Springer, 2009: 831–838.
- [11] JOBE J M, POKOJOVY M. A cluster-based outlier detection scheme for multivariate data [J]. Journal of the American Statistical Association, 2015, 110(512): 1543–1551.
- [12] ZHAO X, ZHANG J, QIN X. LOMA: a local outlier mining algorithm based on attribute relevance analysis [J]. Expert Systems with Applications, 2017, 84: 272–280.
- [13] SATHE S, AGGARWAL C C. Subspace outlier detection in linear time with randomized hashing [C]//Proceedings of 2016 IEEE 16th international conference on data mining (ICDM). Barcelona; IEEE, 2016: 459–468.
- [14] LIU H, LI E, LIU X, et al. Anomaly detection with kernel preserving embedding [J]. ACM Transactions on Knowledge Discovery from Data, 2021, 15(5): 1–18.
- [15] MA X, ZHANG X, TANG X, et al. Hyperspectral anomaly detection based on low-rank representation with data-driven projection and dictionary construction [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 2226–2239.
- [16] ZHANG H, YANG J, SHANG F, et al. LRR for subspace segmentation via tractable Schatten-p norm minimization and factorization [J]. IEEE Transactions on Cybernetics, 2018, 49(5): 1722–1734.
- [17] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers [C]//Proceedings of the 2000 ACM SIGMOD international conference on management of data. Dallas; ACM, 2000: 93–104.
- [18] KRIEGEL H P, KRÖGER P, SCHUBERT E, et al. Outlier detection in axis-parallel subspaces of high dimensional data [C]//Proceedings of advances in knowledge discovery and data mining, Paris; Springer, 2009: 831–838.
- [19] ALMARDENY Y, BOUJNAH N, CLEARY F. A novel outlier detection method for multivariate data [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(9): 4052–4062.
- [20] LI Z, ZHAO Y, BOTTA N, et al. COPOD: copula-based outlier detection [C]//Proceedings of 2020 IEEE international conference on data mining (ICDM). Sorento; IEEE, 2020: 1118–1123.
- [21] KELLER F, MULLER E, BOHM K. HiCS: high contrast subspaces for density-based outlier ranking [C]//Proceedings of 2012 IEEE 28th international conference on data engineering. Arlington; IEEE, 2012: 1037–1048.
- [22] ABDULGHAFOOR S A, MOHAMED L A. A local density-based outlier detection method for high dimension data [J]. International Journal of Nonlinear Analysis and Applications, 2022, 13(1): 1683–1699.